DOCUMENT RESUME

ED 074 085 TM 002 447

AUTHOR	Proger, Barton B.; And Others
TITLE	Unequal Cell Frequencies in Analysis of Variance: A Review and Extension of Methodology for Multiple Missing Observations.
SPONS AGENCY	Montgomery County Schools, King of Prussia, Pa.; Pennsylvania Resources and Information Center for Special Education (PRISE), King of Prussia, Pa.; Research and Information Services for Education, King of Prussia, Pa.
PUB DATE	[72]
GRANT	OEG-1-67-3010-2696
NOTE	28p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Analysis of Variance; Bibliographic Citations; *Data Analysis; *Educational Research; *Research Methodology; Research Reviews (Publications); Statistical Data

ABSTRACT

Many researchers assume that unequal cell frequencies in analysis of variance (ANOVA) designs result from poor planning. However, there are several valid reasons why one might have to analyze an unequal-n data matrix. The present study reviewed four categories of methods for treating unequal-n matrices by ANOVA: (a) unaltered data (least-squares solution and unweighted means solution); (b) data substitution (grand mean method, cell mean method, Winer method, Snedecor-Cochran method); (c) data deletion, and (d) data clustering (unreplicated cell mean method, unreplicated random data clustering method, replicated random data clustering method). The methods were compared empirically and theoretical problems with each were discussed. (Author)

TITLE: Unequal Cell Frequencies in Analysis of Variance: A Review and Extension of Methodology for Multiple Missing Observations

AUTHORS: Barton B. Proger, Raymond G. Taylor, Jr.
Paul A. Green, Lawrence H. Cross, Lester Mann

Pennsylvania Resources and Information Center
for Special Education
King of Prussia, Pennsylvania

and

John R. McGowan

Southern Connecticutt State College
New Haven, Connecticutt

CORRESPONDENCE: Dr. Barton B. Proger
Director of Evaluation and Dissemination

Pennsylvania Resources and Information Center
for Special Education

443 South Gulph Road
King of Prussia, Pennsylvania, 19406

(ABSTRACT)

UNEQUAL CELL FREQUENCIES IN ANALYSIS OF VARIANCE: A REVIEW AND EXTENSION OF METHODOLOGY FOR MULTIPLE MISSING OBSERVATIONS

Many researchers assume that unequal cell frequencies in analysis of variance (ANOVA) designs result from poor planning. However, there are several valid reasons why one might have to analyze an unequal-n data matrix. The present study reviewed four categories of methods for treating unequal-n matrices by ANOVA: (a) unaltered data (least-squares solution and unweighted means solution); (b) data substitution (grand mean method, cell mean method, Winer method, Snedecor-Cochran method); (c) data deletion, and (d) data clustering (unreplicated cell mean method, unreplicated random data clustering method, replicated random data clustering method). The methods were compared empirically and theoretical problems with each were discussed.

(COMPLETE TEXT)

UNEQUAL CELL FREQUENCIES IN ANALYSIS OF VARIANCE: A REVIEW AND EXTENSION OF METHODOLOGY FOR MULTIPLE MISSING OBSERVATIONS[1]

Introduction

The majority of experimental studies in educational research that concern the analysis of variance (ANOVA) contain equal cell frequencies. Since most of these investigations are completed in tightly controlled university settings or laboratory situations, it is almost always possible to ensure that sufficient Ss are available to produce an equal-n data matrix in a factorial ANOVA design. Thus, it is not surprising that most commonly used texts in educational statistics discuss only the equal-n, factorial ANOVA solution. Further, many applied statisticians take the attitude that a researcher has done poor pre-experiment planning if he allows himself to get into an unequal-n circumstance; one is even made to feel guilty about it!

Unfortunately, these equal-n biases of the majority of educational statisticians do little for the researchers in large public school situations where unequal-n ANOVA problems are the rule rather than the exception. Apart from possible lack of adequate planning for the experiment, what are some common reasons for unequal n's to arise in the factorial design? One important reason is inherent dearth of some types of Ss; this consideration is especially prominent in the study of various handicapped populations. If one wants to include such types of Ss in his study, he either must balance them with a like number (pitifully small) of other groups for his study, or he must settle for an unequal-n design. A second reason might be inadvertent experimental mortality (attrition) over the course of the experiment, where one would not for some reason have enough supplementary Ss to substitute for the missing ones in the data matrix. A third reason could be forced experimental mortality during the study when the investigator learns that some of his Ss who had previously been identified as being appropriate to the study, really are not suitable; thus, rather than discard the whole study, the experimenter analyzes his remaining unequal-n matrix. However, whatever the reasons for attempting to analyze an unequal-n data matrix, the range of methods for treating such matrices are relatively unfamiliar to most researchers. The purpose of this paper is to survey existing methods of both common and out-of-the-way nature, as well as to introduce some previously unpublished techniques.

PROCEDURE

Data:

To facilitate discussion of the methods described herein and to provide

readers with a means of verifying the accuracy of their understanding of the analytical techniques, an empirical comparison of all procedures was undertaken by means of one master data matrix for a 3x3 design. Winer (1962) made an initial step in this direction when he used empirical comparisons between least-squares and unweighted-means ANOVA; the present study extends the empirical comparison notion by also including 7 other unequal-n techniques, as well as the original equal-n solution. Table 1 shows an equal-n matrix where the hypothetical investigator intended 15

Insert Table 1 about here

---

independent observations to be contained in each cell. The matrix reflects a typical unequal-n situation often occurring in the remediation of mentally handicapped children where one applies treatments (Factor B). In particular, the hypothetical example assumes that 3 perceptual-motor training programs (the worst being A_1, A_2average, and A_3 best) were given to 3 levels of intelligence (the range of B_1 being 91-105; B_2, 76-90; B_3, 61-75). The criterion is assumed to be the visual sequential memory subtest of the Illinois Test of Psycholinguistic Abilities (ITPA), with a possible score range of 0 to 41. The Os are assumed to be of chronological age 6 to 8 years. The data generation for this empirical simulation was aimed at producing quite strong main effects for factors A and B but quite negligible interaction between the two. Further, to achieve the common happening in which, regardless of mean differences among factorial levels, score ranges across cell categories often overlap to a certain extent, the ITPA scores were allowed to telescope as shown in Table 2. The degree of overlap is consistent across levels within

Insert Table 2 about here

either factor. The individual scores in each cell of Table 1 were generated by 9 independent randomizations based upon the range limits set in Table 2 (Rand Corporation, 1955). The complete 10/cell-n matrix was used only as a pivot for discussion in comparing the several unequal-n procedures. Each unequal-n analysis was computed on the data matrix that results from Table 1 when the italicized entries were deleted. For the unequal-n matrix derived from Table 1, one sees that the cell frequencies range from 10 to 15, with no proportionality among rows or columns assumed; that is, the unequal-n matrix in this study is the "worst" that could arise with respect to the orthogonality issue.

Analyses: Since the majority of unequal-n techniques are not available in programmed form, all computations were completed by electronic calculator, with systematic checking to ensure accuracy. A total of 10 unequal-n procedures were compared in this study. A procedure is described at length only if it is not available elsewhere. The 10 methods can be grouped under four major headings.

(1) UNALTERED DATA: The two unequal-n techniques that fall under this heading are also the most widely known, used, and programmed approaches out of the 10 discussed in this paper. The two methods are known as least-squares analysis and unweighted-means analysis. As pointed out by Winer (1962), in cases where the levels of one factor are proportional to actual population strata so that irregular cell frequencies result naturally, then least-squares ANOVA is appropriate. However, if unequal frequencies in the resultant working sample are not related to the population in a natural proportionality (that is, unequal cell frequencies might be the result of random attrition), then unweighted-means ANOVA is better suited to unequal cell frequencies.

Perhaps the best account of least-squares ANOVA is given by Winer (1962, pp. 224-227, 291-297). Other readable accounts can be found in Snedecor & Cochran (1967, pp. 477- 83, 488-493) and in Ferguson (1966, pp. 319-323). For those particularly interested in trend ANOVA, one should consult Gaito (1965), Black & Davis (1966), and Ferguson (1966, pp. 343-346). For further reading, see Kempthorne (1952, pp.80-81), Rao (1952, pp. 96-98), Gourlay (1955), Snedecor (1956), Kenney & Keeping (1954), Wilk & Kempthorne (1956), Brandt (1932), Strand & Jessen (1943), Yates (1934), Stevens (1948), and Federer & Zelen (1966).

When circumstances behind an unequal-n data matrix indicate that unweighted-means ANOVA is appropriate, one can refer to the examples given in Winer (1962, pp. 103-104, 222-224, 241-244, 374-378) and Snedecor & Cochran (1967, pp. 475-477). For further reading, see Gowen (1952).

(2) DATA SUBSTITUTION: Four methods are worthy of consideration: (a) substitution of the grand mean, (b) substitution of the cell mean, (c) substitution alà Winer, and (d) substitution alà Snedecor & Cochran. All four procedures have in common the attempt to add bits of data to the original unequal-n matrix until it becomes, literally, an equal-n paradigm amenable to classical ANOVA. The only modifications that must be made to the classical statistical machinery is, logically enough, to adjust the degrees of freedom for both within-cells variation and total variance.

For the grand mean method, the mean of the entire unequal-n matrix is computed and substituted for each bit of missing data. For the cell mean method, wherever a cell has one or more missing values, the mean of that cell is computed and substituted for each missing score within that cell.

The substitution method of Winer (1962, p. 281) was designed for situations in which an entire cell is missing! However, in most real-life unequal-n matrices, one almost always has some data within every cell. Thus, the logical extension of Winer's method in which one obtains row (and column) means of the cell means within the row (and column) that contains the missing cell, is to obtain comparable row (and column) means using every individual child's score (including the scores in the deficient cell).

Further discussions of data substitution can be found in Cochran and Cox (1957, pp. 80, 110, 125, 227, 302, 400, 413, 450, 512), Healy and Westmacott (1956), Lindquist (1953, p. 148), Afifi and Elashoff (1966), Lord (1955), Federer (1955, pp. 124-127, 133-134),and Bennett and Franklin (1954, pp. 382-383). Snedecor and Cochran (1967, pp. 320-321) and Li (1964, pp. 231-236-237) present a very interesting iterative procedure for supplying two or more missing values in the data matrix. Basically, one chooses any one of the two or more missing values, estimates a reasonable value, and makes the substitution. The other missing value is estimated with a least-squares formula as though there were only one value missing. Then one goes back and estimates the first value on the basis of the second one and so on, back and forth, until the values change only by very small amounts. Degrees of freedom are again adjusted for total sum of squares and error sum of squares after stabilization has occurred. For exact least-squares methods of data substitution, see Li (1964, pp. 227-243). Winer (1962, pp. 281-283) also provides a method that minimizes the interaction effects. Another basic reference with examples is Snedecor and Cochran (1967, pp. 317-321). Finally,

Examples of data substitution can be readily found in special education research (e.g., Bloom, 1967; Prehm, 1967; Halpern, Mathieu, & Butler, 1968).

(3) DATA DELETION: Another major attempt to form an equal-n matrix from an originally unequal-n paradigm is to use random deletion of cell entries. One looks at the n of the smallest cell and accordingly "prunes" all other cells down to that size. Independent runs through tables of random numbers are used to accomplish an unbiased deletion in the "oversized" cells.

Closely related to the topic of random deletion of observations is the systematic deletion of highly discrepant observations. Snedecor and Cochran (1967, pp. 321-323) present a very enlightening discussion on the rejection of extreme observations. Most rejection methods are based on tests of significance of residuals of observations from expected values. Edwards (1960, pp. 166-168) also describes a method for rejection of discrepant observations on the basis of confidence intervals. Mainland (1968), on the other hand, takes opposition to all methods of rejecting observations; the reader is advised to examine Mainland's notes before employing test-of-significance methods. For further reading, see Anscombe (1960), Anscombe and Tudey (1963), Li (1964, pp. 239-240), and Searls (1963). Some interesting examples of data deletion in applied situations are Shubert, Jansen, & Fulton (1967) and Dawson (1967).

(4) DATA CLUSTERING: In line with the philosophy of the attempts of data deletion and data substitution to form equal-n matrices out of unequal-n ones, the data clustering techniques coalesce several observations within a cell into fewer observations but with no loss or gain in data. The data

clustering techniques are without doubt the least known of unequal-n methods; indeed, some of the procedures to be described here have never been published before.

The only data clustering technique that has been discussed at all is ANOVA where cell means become the units of analysis. In data matrices where all cells have some entries, but cell discrepancies are such as to violate the approximately-equal frequency rule, the within-cells variation is ignored. The highest-order interaction is used as the estimate of error; however, the assumption must be made that the interaction is negligible. In effect, the ANOVA is carried out as though single replication were the case. The basic mathematical defense of the method is given by Finney (1960, p. 48) in terms of differential coefficients of regression functions. The use of interactions as error terms is discussed by Edwards (1960, p. 211), Ferguson (1966, pp. 310-311, 314-316), Lindquist (1953, p. 114), and Scheffe (1959, pp. 247-146). An example of using the highest-order interaction as error is given by Ling (1968).

A new procedure of random data clustering was devised in late 1968 or early 1969 by J. R. McGowan but never before published.[4] He suggested forming random clusters of data within each cell of the original unequal-n matrix. The number of randomly formed clusters is the same as the number of original entries in the smallest cell. In this sense, the method might be called unreplicated random data clustering because some of the clusters will never have more than 1 observation. For example, if the smallest cell has two entries, then in a cell with seven entries, four data would be randomly assigned to one cluster and the remaining three data in that cell would become the second cluster of the cell. Clearly, the clusters in the smallest

cell would always contain only one score each. In the example cited, each cell would contain two clusters, each cluster in turn holding varying numbers of data. After randomly assigning within a cell all original scores to their new cluster "identities", the average of each within-cell cluster is computed. The resulting matrix of equal-frequency, mean data is subjected to a regular equal-frequency ANOVA with the new number of averages taken as the number of data. As far as the authors know, McGowan was the first to put forth such a method. The technique seems to hold interesting possibilities. It should be noted that if the smallest cell has only one original observation, then the "random cluster" method becomes merely cell-means ANOVA (single replication), mentioned just above. In the present example, cell $A_1 B_1$ is of size 13, while the smallest size of any cell is 10. One wants 10 clusters per cell. The only combination of double clusters (those with 2 scores) and single clusters (those with only 1 score) that yield a total of 10 clusters and still use all 13 individual scores, is 3 doubles and 7 singles. To determine which observations within all $A_1 B_1$ go into which of the double and single clusters, the cluster numbers (labels) of 0 to 9 are assigned from a table of random numbers to the observations in the order that the latter are listed within the unequal-n data matrix. Once a digit occurs the second time, it cannot be used again. Further, since one wants only 3 double clusters, only 3 of the digits can be allowed to occur the second time. The averages of all double clusters are computed and, along with the single clusters of the original observations, are entered into a new equal-n matrix upon which the classical ANOVA is finally computed.

The last method compared in this study is an extension of the preceding

clustering technique and might be termed replic ted random data clustering; that is, no cluster will ever have fewer than 2 observations. In the present example where the smallest cell size is 10, one wants to generate 5 clusters in each cell so that at least 2 observations per cluster result.

RESULTS AND DISCUSSION

The summary ANOVA table for 9 unequal-n methods are presented in Table 3. While it must be remembered that the results are only an empirical comparison

---

Insert Table 3 about here

---

within a limited numerical example, one can draw some conclusions. First, one needs some basis for comparison before he can suggest that a certain unequal-n method appears to be a rather poor or good approximation to what would have been the results of the original equal-n experiment. Since the data in this illustration were quite carefully selected to reflect pre-specified differences and to avoid unwanted biases, the complete equal-n solution was available to serve as the basic "control" analysis. One can see the strength of the two main effects, the negligibility of the interaction, and the relatively small within-cells variation. Because the equal-n solution would normally be unavailable, the exact least-squares ANOVA is perhaps the most appropriate "control" for all other unequal-n methods to be compared with. Even though the random attrition of the hypothetical example would dictate the unweighted-means solution, least-squares ANOVA is a better approximation.

The most discrepant set of results occurs in connection with data substitution by the grand mean. Where there should have been a quite negligible interaction, a significant one emerged. On the other hand, substitution by

cell means is a quite accurate approximation of the equal-n results.

When one turns to theoretical considerations of the separate unequal-n methods, a number of interesting insights are yielded. First, one returns to the notion that unequal-n designs can be avoided by sound pre-experiment planning. When one considers an area such as handicapped children (special education), most research does not yield equal cell frequencies. It is difficult enough to get equal numbers of, say, educable mentally retarded Ss for various treatments to be compared on just the factor of treatments itself, but even more difficult to get an equal distribution of sex within the equi-sized EMR groups under each treatment to produce a factorial design. Adding more control variables usually leads to even greater fluctuations in cell frequencies. Thus special education researchers seem more content to measure differences only among treatments in nonfactorial, one-way designs. When an investigator uses one-way ANOVA, valuable information on interactions with non-treatment variables (such as sex, age-level, level of previous functioning, class of brain damage, etc.) is lost.

Nonetheless, proper pre-experiment planning should not be dismissed lightly with regard to avoiding unequal-n data matrices. Consider the case of a three-way factorial ANOVA design in which the factors are treatments, sex, and levels of auditory impairment. A control variable such as auditory impairment that lends itself to a numerical continuum often leads to unequal cell frequencies when the design paradigm is further subdivided by other control variables, such as sex. In the present example, during the planning stages of the experiment, auditory impairment of all potential candidates for participation in the study is determined. A stratification problem, inherent in control variables of continuous type, is then posed. The researcher must

decide whether he want to form control strata on the basis of realistic special education criteria or on the basis of computational expediency. On the latter case, equal cell frequencies can be established no matter how artificial the cut-off points. Too often both theoretical and applied statisticians get side-tracked in trying to establish perfect designs and avoiding statistically difficult, but perhaps more meaningful and generalizable, situations. Of course, even if artificial stratification points have been chosen on the control variable distributions for achieving equal cell frequencies, experimental attrition may occur during the experimental period. For further reading, see Hess, Sethi, & Balakrishnan (1966).

However, even the best of experimental planners cannot avoid every pot-hole in the road of design. Consequently, statistical methods for handling unequal frequencies must be considered. With regard to the first category of unequal-n methods (those dealing with unaltered data), Winer (1962) claims least-squares ANOVA provides more powerful tests of significance than unweighted-means solutions. It should be cautioned that one basic difference between least-squares ANOVA and unweighted means ANOVA is that the variance relation among the total, between, and within components holds only for the least-squares method. In other words, true orthogonality of variance components exists only for the least-squares ANOVA. The only apparent difference between least-squares ANOVA and unweighted means ANOVA is in obtaining a best-fit regression model based on cell means and average frequencies without response surface regression weighting. Basically, in a least-squares two-way ANOVA, one solves a set of normal equations analogous to that in multiple regression. As in covariance analysis, one makes adjustments to the raw sums of squares. He uses the exact cell, column and total frequencies

along with cell totals. First, one computes unadjusted row, column, and cell sums of squares. There are then two options: (a) $SS_{ab(adj.)}$ can be computed directly from means of cell means or, (b) one can go through the unadjusted, exact frequency analysis, computing $SS_{b\ (adj.)}$ and $SS_{a\ (adj.)}$ by the abbreviated Doolittle Algorithm or, somewhat easier, by the Dwyer square-root algorithm, and then obtain $SS_{ab(adj.)}$ by subtraction. To use a physical analogy, if one pictures different thickness poker chips for different magnitude scores arranged vertically one on top of the other in their respective cells, the least-squares ANOVA drops a response surface blanket over the stacks of chips naturally, taking into account different frequencies as well as different sizes of scores. On the other hand, unweighted-means ANOVA does not throw the blanket down over what exists; rather, it statistically builds by leveling off the peaks and then fits a uniform unweighted surface on the situation, taking account only of differences in cell score averages.

In dealing with least-squares solutions, an important and generally unappreciated issue is that of how far the observed frequencies can deviate from the frequencies expected under proportionality. This question could be attacked by an application of factorial Chi-square analysis. However, since Chi-square is a test of poor power, its results cannot be relied upon too heavily. The present authors contend that least-squares and unweighted-means ANOVA are applied too often in situations where their mathematical appropriateness cannot be justified. This is especially unfortunate because the tests of appropriateness are themselves rather weak and under-powered. Under expected equal frequencies, Snedecor and Cochran (1967) suggest that discrepancies in cell frequencies should lie within a 2 to 1 ratio, but only if the majority of cell frequencies are in closer agreement. However, this rule is given

without any mathematical evidence to support it. Ferguson (1966, pp. 319-323) provides a discussion of ANOVA from the standpoints of Tsao's (1946) methods for equal and proportional expected frequencies. However, the reader must be aware of the possibility of bias, both positive and negative, in F tests when deviations from the expected frequencies are large. Unfortunately, one has no completely satisfactory method of testing such deviations. Similarly, turning from least-squares solutions to unweighted-means techniques, one worries about how much variation can be allowed among the unequal-n's relative to the original expected frequencies. The situation is compounded by the fact that one uses the harmonic mean of the observed cell frequencies in obtaining sums of squares, rather than the original frequencies.

The second major set of unequal-n methods deals with substitution of data to obtain an equal-n matrix. Beginning with the grand mean method, one might suspect that it would produce a very poor approximation to the original unequal-n matrix, or at worst, to the least-squares unequal-n solution. The fact that the grand mean probably is not really close to any specific cell means distorts the original cell means quite a bit, as well as increasing within-cells variation.

More positive things can be said about the second technique of data substitution: insertion of cell means for a cell's missing observations. First, substitution of cell means does not change the original cell mean. Second, and perhaps most importantly, the method does not affect the within-cells variation. Finally, the technique provides a very good approximation to both the least-squares and equal-n solutions.

The data substitution method of Winer, as modified for purposes of this paper, makes use of both main effect means and the grand mean. The basic structure underlying Winer's technique is both logical and pleasing

in its ease of application. However, the method suffers from some severe limitations: (a) it assumes no interactions of any significance; (b) generalized, "halo" distortion occurs simply because of data from outside the cell of interest entering into the estimation; and (c) severe distortion occurs if the cell with missing data lies at either end of a score continuum. Winer suggests that a preferable alternative would be to use a multiple regression equation in connection with the response surface of the experiment.

Both the original Winer and Snedecor-Cochran substitution methods were designed for cells that had no data at all in them. While Winer's method could be modified to allow any data that might be available within the cell of interest to enter into the substituted data estimates, the method of Snedecor and Cochran must remain in its original form and thus could not be used in the empirical comparison of this study.

Some final comments on data substitution are in order. In realistic learning situations where it is likely that experimental mortality will occur in a one-day study, the investigator might consider running a separate replication of the primary study so as to have data in reserve for substitution purposes. It seems statistically more pleasing to substitute real data than to make elaborate assumptions about the response surface. For example, if the desired cell size is 5, and if one cell is missing 2 observations for purposes unrelated to the experiment, then the corresponding data cell from the reserve replication would be randomly "robbed" of 2 entries. The cautious researcher would then reduce the degrees of freedom for both the error and total sums of squares by 2. Of course, in any method of data substitution, the degrees of freedom for the error sum of squares and the total sum of squares have to be adjusted accordingly; clearly, the principle of

diminishing returns applies, since the error mean square becomes larger in the process.

Both conceptually and practically, data deletion, the third major category of methods for treating unequal-n data matrices, seems quite weak. The technique is suitable only when the original cell size expectation is large. This procedure can be extremely wasteful if cell frequencies are highly discrepant. A workable compromise is to find the optimum combinations of data substitution and data deletion in order to achieve the least amount of "synthetic" data in balance with the maximum degrees of freedom. Whether or not a subject is to be discarded from analysis is an issue which only the investigator can decide. However, leaving all original data present and unmodified seems to be the most defensible course. Suppose, for example, that a normal pupil refused to cooperate on a test or was obviously working far below his level. Many analysts would either discard this data or at least regression-modify it. Clearly, these procedures violate reality. If normal pupils occasionally behave erratically, then the analysis should reflect this fact, not ignore it.

The last group of unequal-n data techniques concern data clustering. The use of original cell means as the unit of analysis is the only familiar method of clustering; in other words, one has turned his unequal-n data matrix into an equal-n, single replication design. There is very little in the literature about single replication studies where all factors are fixed. Ferguson (1966, p. 311) discusses this situation briefly. Perhaps one could reason that, if the highest-order interaction of completely fixed factors is to be the error term, or at least part of it, then (since this "error" is not operating in a random fashion) it would comprise a systematic overestimate. In this case, a randomly operating error term is treated as though

it is the minimum error that could exist, and the more systematic the error term, the greater the inflation. In other words, a non-negligible interaction chosen as an error term can be considered an upper limit to the error. At worst, one has conservative tests of his main effects and other interactions.

To what extent must the assumptions of homogeneity of variance and normality be met in the case where cell means are used as the units of analysis? There is zero variability within each cell. Normality of individual scores cannot even be considered. Perhaps these thoughts, along with the robustness of the F test, make this method of analysis one of the soundest of all. However, one should not assume that ANOVA by cell means is foolproof; Finney (1960, pp. 88-89) considers the procedure appropos only when the design is "saturated" with factors, say 6 or more. For further discussions about violation of basic ANOVA assumptions, see Snedecor and Cochran (1967, pp. 278, 324-325), Scheffe (1959, pp. 360-364), Edwards (1960, pp. 125-128, 132), Box (1953), Box (1954), and Lindquist (1953, pp. 72-90).[5]

The other two methods of data clustering (replicated and unreplicated random data clustering) appear pleasing at first glance because they retain all original bits of the unequal-n data, do not substitute contrived and distorted data, and yield equal-n's for classical ANOVA to be applied. Further, the replicated version seemed to offer somewhat greater reliability of individual cluster means than the unreplicated technique. In spite of these apparent advantages, the empirical comparison demonstrated that both techniques were poor approximations to the equal-n and unequal-n control solutions.

SUMMARY

This paper has brought together within a single perspective several distinct methods for handling complicated, unequal-n data matrices in ANOVA. A discussion of each technique's virtues and problems was presented. Further, an empirical comparison within a tightly controlled numerical example was undertaken among the methods. Substitution by cell means appeared to give the most accurate approximation to the original equal-n solution, as well as to the least-squares unequal-n results. However, in the final analysis, only formal mathematical statistics can establish the superiority of one method over the other. It is hoped this paper will give impetus to mathematical research into the relative theoretical properties of each technique.

The investigators wish to conclude the review by cautioning the reader to be thoroughly familiar with the limitations placed upon each method; none of the techniques presented are "foolproof." No one method suffices for every unequal-n problem the applied researcher meets from day to day. Some procedures have more severe restrictions than others. With some thought, the reader can devise completely new techniques, as well as modifications of those presented in this paper. The field of unequal-n ANOVA methodology is far from being a "dead" research topic in applied statistics. It should be noted, however, that some statisticians would disapprove of several of the methods discussed here, if for no other reasons on philosophical grounds.

In conclusion, it would be nice if the investigators could tell the readers to use computer programs for all their unequal cell-frequency

needs. This cannot be done. While several programs do exist, it must again be emphasized that most are appropriate only for certain situations. Many of the more refined programs are difficult to use, and several have such poor documentation of computational procedure that the user does not know which of the methods surveyed in this review he is using.

FOOTNOTES

[1]The writing of this paper was jointly supported by Research and Information Services for Education (RISE) under Title III of the Elementary and Secondary Education Act of 1965 (OEG-1-67-3010-2696); by Pennsylvania Resources and Information Center for Special Education (PRISE), also under Title III (R-22-H, 48-70-0003-0); and by Montgomery County Intermediate Unit No. 23. However, the opinions expressed herein are solely those of the investigators and do not necessarily reflect the position or policy of the supporting agencies. BBP is responsible for the review of literature and for the conceptualization of the different methods of treating unequal-n data matrices. JRMcM provided the basic idea behind the data-clustering techniques, as well as valuable criticism of the basic thinking in this paper. RGT and LM also aided in conceptual criticism. Finally, PAG and LHC performed the empirical analyses for this study.

[2]The investigators welcome correspondence relating to this article. Address all comments to Dr. Barton B. Proger, Director of Evaluation and Dissemination, Pennsylvania Resources and Information Center for Special Education, 443 South Gulph Road, King of Prussia, Pennsyivania 19406.

[3]Some of the mathematical premises behind estimation of missing values by minimization of residual sum of squares have been discussed by Jaech (1966) and by Sclove (1972) and have subsequently been commented upon in miscellaneous "letters to the editor" on pp. 57-58 in The American Statistician for October, 1972.

[4]Alass, Peckham, and Sanders (1972) studied violation of basic ANOVA assumptions (non-independence of errors, non-normality, and heterogeneous variances) for both equal-n matrices and unequal-n matrices. However, the investigators in that study were not interested per se in different methods of treating unequal-n data matrices.

PRIMARY REFERENCES

BLACK, Alan H. and Davis, Leo, Jr., "The relationship between intelligence and sensorimotor proficiency in retardates" American Journal of Mental Deficiency, 71(1), (July 1966), pp. 55-59.

BLAKE, Kathryn, A., Ainsworth, Stanley H., and Williams, Charlotte L. "Effects of induction and deduction on deaf and hearing individuals' attainment of first-order concepts" American Annals of the Deaf, 112, (Sept. 1967), pp. 606-613.

BLOOM, Wallace "Effectiveness of a cooperative special education vocational rehabilitation program" American Journal of Mental Deficiency, 72, (Nov. 1967), pp. 393-403.

DAWSON, William W. "Stereoscopic perceptual deficit of mentally retarded adults" American Journal of Mental Deficiency, 71(6), (May 1967), pp. 964-968.

EDWARDS, A. L. Experimental design in psychological research. Rev. ed., New York: Holt, Rinehart & Winston, (1960).

FERGUSON, G. A. Statistical analysis in psychology and education. 2nd ed., New York: McGraw-Hill, (1966).

FINNEY, D. J. An introduction to the theory of experimental design. Chicago: University of Chicago Press, (1960).

FOULKE, Emerson "Comparison of comprehension of two forms of compressed speech" Exceptional Children, 33(3), (Nov. 1966), pp. 169-173.

HESS, V.K., Sethi, V. K., & Balakrishnan, T. R. "Stratification: A practical investigation" Journal of American Statistical Association, 61, (1966), pp. 74-90.

HUDDLE, Donald D. "Work performance of trainable adults as influenced by competition, cooperation, and monetary reward" American Journal of Mental Deficiency, 72, (Sept. 1967), pp. 198-211.

KAHN, Harris, and Burdett, Arthur D. "Interaction of practice and rewards on motor performance of adolescent mental retardates" American Journal of Mental Deficiency, 72, (Nov. 1967), pp. 422-427.

LEE, Marilyn C. "Interactions, configurations, and nonadditive models" Educational and Psychological Measurement, 21(4), (Winter 1961), pp. 797-805.

Li, C. C. Introduction to experimental statistics. New York: McGraw-Hill, (1964).

LINDQUIST, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton-Mifflin, (1953).

LING, Daniel "Three experiments on frequency transposition" American Annals of the Deaf, 113(2), (March 1968), pp. 283-294.

LLOYD, Lyle L., Rolland, John C., & McManis, Donald L. "Performance of hearing impaired and normal hearing retardates on selected language measures" American Journal of Mental Deficiency, 71(6), (May 1967), pp. 904-908.

LUBAN, Ardie "The interpretation of significant interaction" Educational and Psychological Measurement, 21(4), (1961), pp. 807-817.

MAINLAND, D. Notes on biometry in medical research. VA Monograph 10-1 Supplement 4, (August 1968), Washington: Veterans Administration.

MAYER, C. Lamar "The relationship of early special class placement and the self-concepts of mentally handicapped children" Exceptional Children, 33, (Oct. 1966), pp. 77-81.

PREHM, Herbert J. "Studies in paired associates learning: 1. An examination of methodology" American Journal of Mental Deficiency, 72, (Nov. 1967), pp. 492-495.

SCHEFFÉ, H. The analysis of variance.. New York: Wiley, (1959).

SHUBERT, O. Wendell, Jansen, Barbara C., & Fulton, Robert T. "Effects of speech improvement on articulatory skills in institutionalized retardates: II" American Journal of Mental Deficiency, 72, (Sept. 1967), pp. 212-214.

SNEDECOR, G. W., & Cochran, W. G. Statistical methods. 6th Ed., Ames (Iowa): Iowa State University Press, (1967).

STANLEY, Julian C. "Introduction: Symposium: Interactions in psychometrics and experimentation" Educational and Psychological Measurement, 21(4), (Winter 1961), pp. 793-795.

WINER, B. J. Statistical principles in experimental design. New York: McGraw-Hill, (1962).

Secondary References

AFIFI, A. A., and Elashoff, R. M. "Missing values in multivariate statistics 1. Review of the literature." Journal of American Statistical Association, 61, (1966), pp. 595-604.

ANSCOMBE, F. J. Technometrics, 2. (1960), p. 123.

ANSCOMBE, F. J., and Turkey, J. W. Technometrics, 5, (1963), p. 141.

BENNETT, C. A., and Franklin, N. L. Statistical analysis in chemistry and the chemical industry, New York: Wiley, (1954).

BINDER, A. "The choice of an error term in analysis of variance designs." Psychometrika, 20, (1955), pp. 29-50.

BOX, G. E. P. "Effects on inequality of variances in the one-way classification." Annals of Mathematical Statistics, 25, (1954), pp. 290-302.

BOX, G. E. P. "Non-normality and tests on variance." Biometrika, 40, (1953), pp. 318-335.

BOZIVICH, H., Bancroft, T. A., and Hartley, H. O. "Power of analysis of variance procedures for certain incompletely specified models." Annals of Mathematical Statistics, 27, (1956), pp. 1017-1043.

BRANDT, A. E. Unpublished doctoral dissertation, Iowa State College. (1932).

BUTLER, R. A. Journal of Exceptional Psychology, 48, (1954), p. 19.

COCHRAN, W. G., and Cox, G. M. Experimental designs. 2nd Ed., New York: Wiley, (1957).

FEDERER, W. T. Experimental design. New York: Macmillan, (1955).

FEDERER, W. T., and Zelen, M. Biometrics, 22, (1966), p. 525.

GOWEN, J. W. American Journal of Human Genetics, 4, (1952), p. 285.

GOURLAY, N. "F-test bias for experimental designs in educational research." Psychometrika, 20, (1955), pp. 227-248.

GREEN, B. F., and Turkey, J. "Complex analysis of variance: General problems." Psychometrika, 25, (1960), pp. 127-152.

HEALY, M., and Westmacott, M. Applied Statistics, 5, (1956), p. 203.

KEMPTHORNE, O. The design and analysis of experiments. New York: Wiley, (1952).

KENNEY, J. F., and Keeping, E. S. Mathematics of statistics. Part 1. 3rd Ed., Princeton: Van Nostrand, (1954).

LORD, F. M. "Estimation of parameters from incomplete data." Journal of American Statistical Association, 50, (1955), pp. 870-876.

OLDS, E. G., Mattson, T. B., and Odeh, R. E. Notes on the use of transformations in the analysis of variance. WADC tech. rep., (1956), 56-308, Wright Air Development Center.

RAO, C. R. Advanced statistical methods in biometric research. New York: Wiley, (1952).

RIDER, P. R., Harter, H. L., and Lum, M. D. An elementary approach to the analysis of variance. WADC tech. rep., (1956), 56-20.

SEARLS, D. T. "On the large observation problem." Unpublished doctoral dissertation, North Carolina State University, (1963).

SNEDECOR, G. W. Statistical methods. 5th Ed., Ames (Iowa): Iowa State College Press, (1956).

STEVENS, W. L. Biometrika, 35, (1948), p. 346.

STRAND, N., and Jessen, R. J. Iowa Agricultural Experimental Statistics Research Bulletin, 315, (1943).

TSAO, F. "General solution of the analysis of variance and covariance in the case of unequal or disproportionate numbers of observations in the subclasses. Psychometrika, 11, (1946), pp. 107-128.

TUKEY, J. W. "One degree of freedom for nonadditivity." Biometrics, 5, (1949), pp. 233-242. Queries, Biometrics, 11, (1955, pp. 111-113.

WILK, M. B., and Kempthorne, O. WADC technical report, Vol. 2, Washington: Office of Technical Services, U. S. Dept. of Commerce, (1956). Report No. 55-244.

YATES, F. Journal of American Statistical Association, 29, (1934), p. 51.

Glass, G. V., Peckham, P. D., & Sanders, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 1972, 42, 237-288.

Jaech, J. L. An alternate approach to missing value estimation. The American Statistician, 1966, 20 (5), 27-29.

Sclove, S. L. On missing value estimation in experimental design models. The American Statistician, 1972, 26 (2), 25-26.

TABLE 1

A_1			A_2			A_3		
19	22	21	23	22	19	28	25	27
22	20	22	27	26	24	30	28	31
14	21	15	24	23	20	32	28	28
18	19	16	22	25	27	32	30	28
22	16	18	19	26	19	25	30	28
14	9	13	21	17	21	22	23	25
12	10	15	14	16	14	23	25	24
10	17	17	14	17	20	27	21	26
16	15	12	22	16	16	25	23	21
14	15	16	15	22	20	27	23	23
9	8	6	16	13	14	15	22	15
11	6	9	14	17	17	16	19	14
10	10	6	15	11	9	15	19	20
6	7	8	10	17	9	14	21	21
5	9	4	17	12	14	14	18	21

TABLE 2

Ranges of Test Scores in Hypothetical Example

Levels of Factor B	Levels of Factor A		
	A_1	A_2	A_3
B_1	14-22	19-27	24-32
B_2	9-17	14-22	19-27
B_3	4-12	9-17	14-22

TABLE 3

	A				B				AB				Within			Total	
	SS	df	MS	F	SS	df	MS	F	SS	df	MS	F	SS	df	MS	SS	df
1. Original Data	2,232.178	2	1,116.089	162.918	2,539.244	2	1,269.622	185.310	28.331	4	7.083	1.034	863.180	126	6.851	5,662.933	134
2. Unweighted Means	2,006.799	2	1,003.399	145.526	1,937.295	2	968.648	140.486	50.494	4	12.623	1.831	751.552	109	6.895	4,747.670	117
3. Least Squares	2,100.478	2	1,050.239	152.401	2,080.512	2	1,040.256	150.952	48.893	4	12.223	1.774	751.152	109	6.891	4,981.035	117
4. Cell Means	156.632	2	78.316	79.469	151.147	2	75.573	76.685	3.942	4	.986	---	---	---	---	311.721	8
5. Random Deletion	1,571.267	2	785.633	108.538	1,365.067	2	682.533	94.295	19.467	4	4.867	.6724	586.300	81	7.238	3,542.100	89
6. McGowan A	1,502.450	2	751.225	132.800	1,525.717	2	762.858	134 857	37.633	4	9.408	1.663	458.200	81	5,657	3,524.000	09
7. McGowan B	783.299	2	391.649	101.640	757.724	2	378.862	98.321	16.483	4	4.121	1.069	138.721	36	3.853	1,696.226	44
8. Data Sub. Grand Mean	1,991.287	2	995.643	108.465	1,944.361	2	972.181	105.909	156.114	4	39.029	4.252	1,000.551	109	9.179	5,092.313	117
9. Data Sub. Winer	2,341.657	2	1,170.828	167.587	2,359.233	2	1,179.616	168.845	38.872	4	9.703	1.389	761.522	109	6.986	5,501.223	117
10. Data Sub. Cell Means	2,349.485	2	1,174.742	170.465	2,267.194	2	1,133.597	164.492	59.131	4	14.783	2.14	751.177	109	6.892	5,426.987	117

TABLE 4

Cell Means of Various Solutions

Cell	Solutions							
	Original equal-n	Original unequal-n	Cell* Mean Sub.	Grand* Mean Sub.	Winer* Sub.	Random Deletion	McGowan A-Cluster	McGowan B-Cluster
$A_1 B_1$	19.000	18.692	18.692	18.610	18.653	18.200	18.692	18.692
$A_2 B_1$	23.067	22.200	22.200	20.825	22.554	22.200	22.200	22.200
$A_3 B_1$	28.667	28.667	----	----	----	27.900	28.667	28.667
$A_1 B_2$	13.667	13.333	13.333	14.282	13.263	13.400	13.333	13.333
$A_2 B_2$	17.667	17.667	----	----	----	17.800	17.667	17.667
$A_3 B_2$	23.867	24.000	24.000	22.420	23.846	23.700	24.000	24.000
$A_1 B_3$	7.600	7.600	----	----	----	7.500	7.600	7.600
$A_2 B_3$	13.667	14.250	14.250	15.015	13.956	14.000	14.250	14.250
$A_3 B_3$	17.600	17.600	----	----	----	18.200	17.600	17.600

*Means calculated by following formula:

$$M = \frac{(n_*)\ (M_u) + (n_+)\ (M_s)}{n_{..}}$$

where $n_{..}$ = original cell size for equal-n (15).

n_* = original cell size for unequal-n.

n_+ = number of data substituted.

M_u = mean of unequal cell data.

M_s = mean of data substituted.